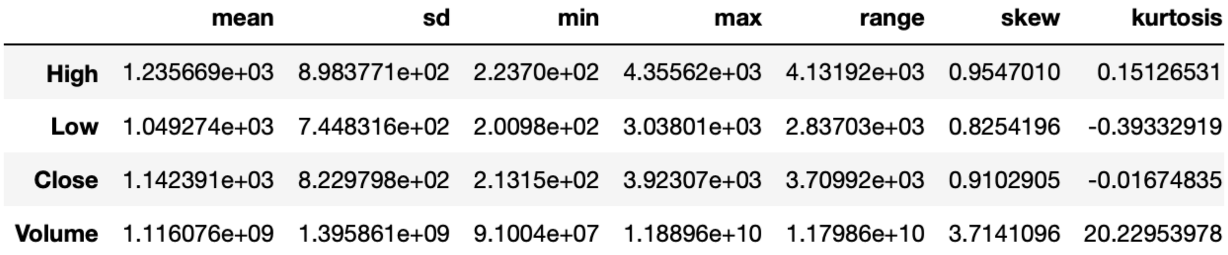
Steve Depp
413 – 55

Assignment 1
1 July 2019

1.1 The following present sample mean, standard deviation, skewness, excess kurtosis, minimum, and maximum for High, Low, Close and Volume of bitcoin prices from 2 Aug 2017 to 20 Feb 2018 (data set provided began in Jul; so this is amended to 2 Aug start):



	mean	sd	min	max	range	skew	kurtosis
High	1.235669e+03	8.983771e+02	2.2370e+02	4.35562e+03	4.13192e+03	0.9547010	0.15126531
Low	1.049274e+03	7.448316e+02	2.0098e+02	3.03801e+03	2.83703e+03	0.8254196	-0.39332919
Close	1.142391e+03	8.229798e+02	2.1315e+02	3.92307e+03	3.70992e+03	0.9102905	-0.01674835
Volume	1.116076e+09	1.395861e+09	9.1004e+07	1.18896e+10	1.17986e+10	3.7141096	20.22953978

Observations:

Overall, prices appear moderately positively skewed while Volume is very positively skewed and very leptokurtotic. All the following comparisons are relative to normal distributions.

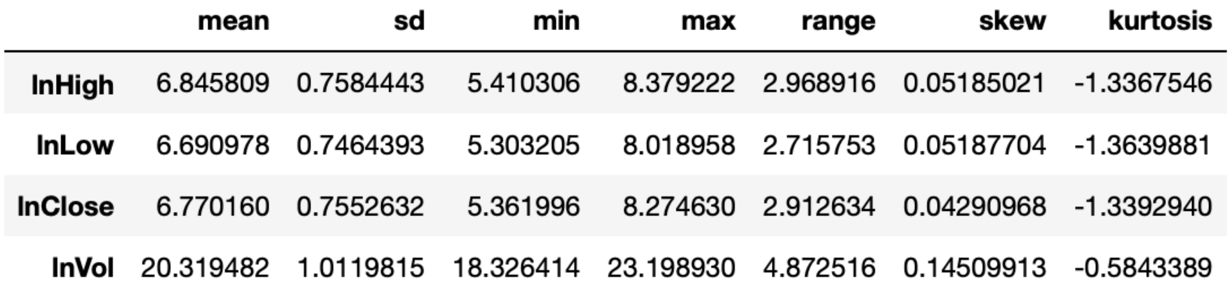
All Close statistics are between the Low and the High, including the mean price as expected.

The standard deviation of High and Close prices is wider than Low which is consistent with the scale of the means and ranges of the distributions. One would need to divide standard deviation into means to get a better feel for relative %spread of High v Low v Close.

All distributions are positively skewed, suggesting observations tails extend to the right more than the left. Volume is highly skewed while the price data is only moderately skewed.

Kurtosis measures excess-kurtosis, and is negative for Low and Close prices and positive for High and very positive for Volume. Positive excess kurtosis suggests a distribution where center and extreme values are more likely than they would be in a normal distribution. This is very much the case for Volume and only moderate for High. Negative excess kurtosis suggests a distribution where shoulder values are more likely than would be seen in normal distribution.

1.2 The following present sample mean, standard deviation, skewness, excess kurtosis, minimum, and maximum for the natural log of High, Low, Close and Volume of bitcoin prices from 2 Aug 2017 to 20 Feb 2018:



	mean	sd	min	max	range	skew	kurtosis
lnHigh	6.845809	0.7584443	5.410306	8.379222	2.968916	0.05185021	-1.3367546
lnLow	6.690978	0.7464393	5.303205	8.018958	2.715753	0.05187704	-1.3639881
lnClose	6.770160	0.7552632	5.361996	8.274630	2.912634	0.04290968	-1.3392940
lnVol	20.319482	1.0119815	18.326414	23.198930	4.872516	0.14509913	-0.5843389

Observations:

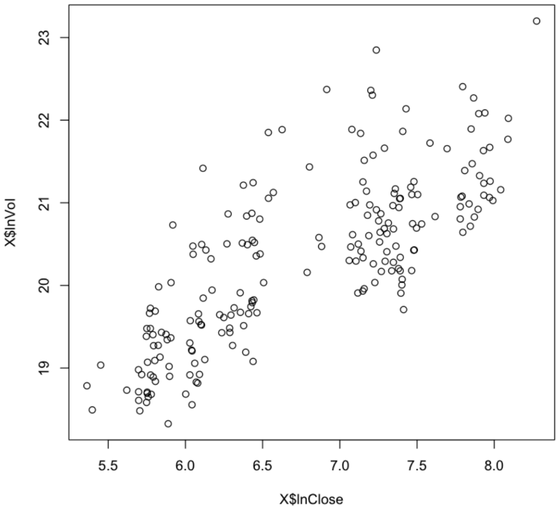
All distributions now show very slight positive excess skew and negative kurtosis.

Skew is much tamed versus the data in (1.1).
Kurtosis is more negative for all data than it was in (1.1);
kurtosis has flipped from positive for High to negative for lnHigh and from very positive for Vol to slight negative for lnVol. In general one might say the log price distributions are now closer to normal in skew terms than price distributions, log price distributions are further from normal in kurtosis terms than price distributions and log Volume is much closer to normal than Volume.
The means and standard deviations of log versions of H L C appear more proximate than simple H L C data, but these are exponents so that is to be expected.

1.3 The sample mean of log Closing prices is 6.770160. We do not know the population variance, and so we use the t-test. As shown below, the t-test yields a test statistic of 127.72 and a p value of less than 2.2e-16 meaning that one can reject the Null hypothesis that the true or population mean log Close price equals zero at the 5% level or with confidence > 99.9%. The p-value suggests the probability of a Type 1 error, rejecting the Null when it is true, is less than 2.2e-16. (More often I would employ this type of test with log returns than log prices as it seems odd to suspect true price = zero.)

```
        One Sample t-test

data:  X$lnClose
t = 127.72, df = 202, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 6.665638 6.874682
sample estimates:
mean of x
  6.77016
```

1.4 The pairwise association between the sample log closing price and log volume in our data can be described in many ways, but succinctly by the correlation coefficient which = 0.7722. This figure represents whether there is a linear relationship between the two sets and usually is not robust to outliers. From the plot below, one can see a clear relationship between the 2 sets which appears to be linear and which appears to be consistent across the range of sample data provided with no outliers.



However, if you examine the LHS versus the RHS you might see on the LHS a slightly higher sloped relationship than the whole sample data and a more amorphous shape on the RHS. Correlation is covariance scaled by individual standard deviations which appear to be wider on the right ¾ of the plot than the left ¼ of the plot; correlation does not see that. Since there is no time component to this plot, and we are asked to consider the entire time frame from Aug to Feb, one cannot say that if you segmented the data in to different time periods or looked and a rolling window of time periods whether there is a stronger association in one time period than another. Lastly if one computes Kendal's tau = 0.5832 you get a sense of concordance versus discordant pairs of observations: when e.g. when 2 pairs of log Close, log Volume observations are compared, when observation A of log Close > observation B of log Close, then observation A of log Volume is > observation B of log Volume, 58% of the time.
As with other statistics reviewed here, we can reject the null hypothesis that the measured correlation and Kendal's tau are zero at the 95% confidence level. Very similar p-values here as above. This does not increase confidence in association, just confidence in measured values are not zero. This t-test employed here is robust to modest departures from normality, but as we will see next, the lnClose prices does not appear normal.

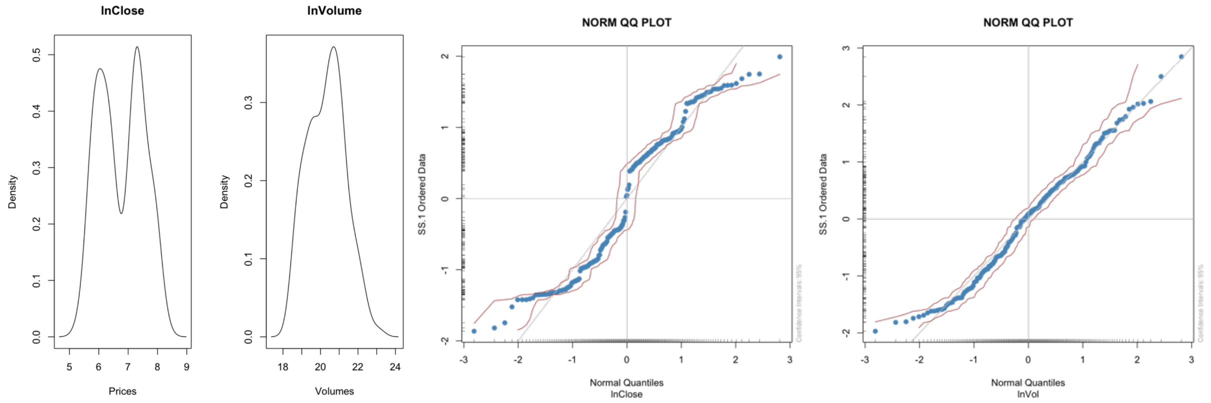
```
        Pearson's product-moment correlation

data:  X$lnClose and X$lnVol
t = 17.231, df = 201, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7099959 0.8224545
sample estimates:
      cor 
0.7722046 


        Kendall's rank correlation tau

data:  X$lnClose and X$lnVol
z = 12.358, p-value < 2.2e-16
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau 
0.5832459 
```

1.5 Density plots and QQ plots

As a whole the lnClose plots lack of conformity to a normal distribution would lead me to question the hypothesis testing conducted with lnClose data since these tests are based on the t-distribution, which is robust to small departures from normality but not to a bimodal departure of this magnitude. The bimodal distribution is consistent with the comment above re separating 2 different relationships between lnClose and lnVolume on the LHS and RHS of the scatter plot.

lnClose
The lnClose bimodal distribution is evident in both QQ plots and density plots. Density dips in the middle around 7 and the QQ plot is vertical in the center of the x axis representing that values which are +/- half a standard deviation in the lnClose data

would appear at zero of a Normal distribution (thus the dip in the density plot). NormQQ demonstrates light tails and the density's 2 modes are effectively heavy shoulders and no peak which is consistent with the negative kurtosis in found in question (1.2). There doesn't appear to be any asymmetry though which is also consistent with (1.2) data.



lnVolume

lnVolume positive skew was large relative to ln price data in (1.2) though not very large absolutely, but one can see some degree of positive skew in the relative positioning of the 2 points on the line in the extreme RHS of the QQ plot versus the 3 points on above the line the extreme LHS of the QQ plot: the LHS is truncated at 2 SD and the RHS extends to 3SD matching the ND. It is difficult to make out the larger LHS shoulder in QQ that is evident in the density plot. So, I am glad we have both.

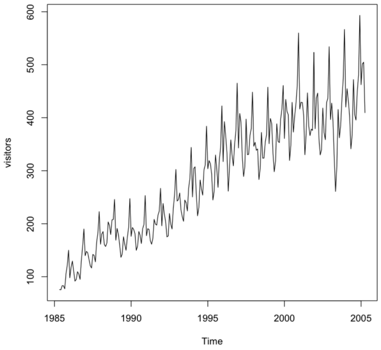
2.1 95% confidence interval for daily log closing price was shown above in (1.3): lower confidence level = 6.665638 and upper confidence level = 6.874682. These figures represent the 95% of the t-distribution of the standard error of the sample mean around the mean of the sample distribution: mean +/- sqrt(var/n) * qt(1-c1/2) with n-1 DOF.

2.2 To test for tail symmetry in the lnClose price set, define null hypothesis H_0: m_3 = zero and alternative hypothesis H_a: m_3 != zero, compute the test statistic = sample skew / sqrt(6/n) which is 0.2496 which is less than 1.96, which means one cannot reject the null hypothesis that the distribution of lnClose is symmetric, skew = m3 = 0.

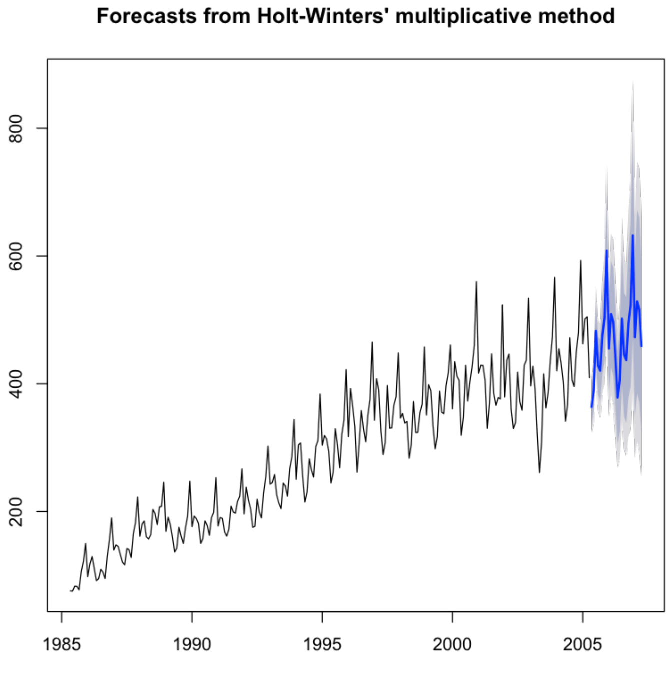
2.3 To test for kurtosis in the lnClose price set, define the null hypothesis H_0: K = 3 and alternative hypothesis H_a: K != 3. This is equivalent to testing the null hypothesis H_0: excess kurtosis = zero and the alternative hypothesis H_a: excess kurtosis != zero. Compute the test statistic = sample excess kurtosis / sqrt(24/n) which is -3.8951 and whose absolute value is greater than 1.96 enabling one to reject the null hypothesis that the distribution of lnClose has non-zero kurtosis.

(Just in case needed, 1.96 is the Z score for the 97.5 percentile of the normal distribution, enabling us to express 95% confidence in a 2 tailed test.)

3.1 Time series plot of Australian overseas visitors. From this plot, it appears there is not only a consistent trend higher in visitors, but and increasing variability in visitors when measured in numbers of persons visiting and possibly reliable seasonality. The span of time covers May 1985 to Apr 2005 or 19 full years in the middle and 2 half years on either end. The plot counts 20 spikes higher over those years. So, there might be seasonal upticks in visitor arrivals consistently found in the data. One can see less pronounced down ticks in visitor arrivals on the LHS and more pronounced down ticks in visitor arrivals on the RHS commencing around 1995. Even as the trend leveled out after the turn of the century, (as the cost of AUD currency reached secular lows), the variation in arrivals becomes increasingly volatile, suggesting a larger percent variation than witnessed prior to 2000. From only this plot of low granularity, it would be difficult to separate variability from reliable seasonality, where one could say that a range of months A to C sees the spikes every year, but the counts to point to some form of pattern.



3.2 Using Holt-Winters multiplicative method.



hw(visitors, seasonal="multiplicative") that produces the above plot is the same as ***hw(visitors, seasonal="multiplicative", trend=”additive”, exponential=FALSE)***. That is, this is Holt’s linear model of additive trend with multiplicative seasonal. Shmueli and Lichtendahl say that “The additive trend model assumes that the level changes from one period to the next by a fixed amount. Hence, the forecasting … adds *k (years of)* trend estimates.” The viability of this method would be difficult to fully assess without competing models, and a validation regimen, but the design of the model does not suit travel arrivals in my opinion. The trend in Holt's linear model is a linear function of h = years. Hyndman notes "The forecasts generated by Holt's linear method display a constant trend (increasing or decreasing) indefinitely into the future. Empirical evidence indicates that these methods tend to over-forecast, especially for longer forecast horizons." That was part of the reason for damped model innovation. With the hefty 1985-1997 trend and what looks to be 2 years of movement higher in the last 2 years, I would rather exponential trends that are proportional to the factored movement of the data and not simply years.

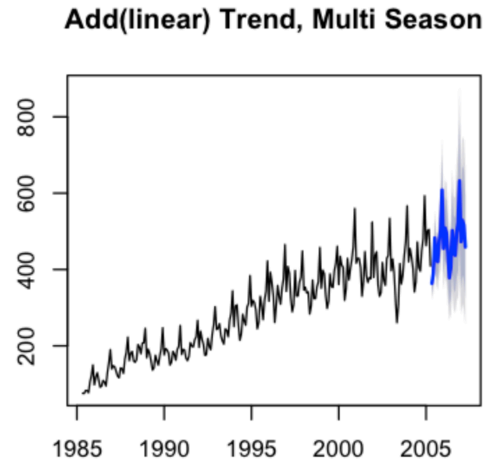
3.3 As with trend, the multiplicative seasonal is necessary because the level of visitors has trended up in the last 2 years (and for longer in the 80s and 90s) and may continue to do so. We would want seasonal variation in the data to increase as the level of the series increases.

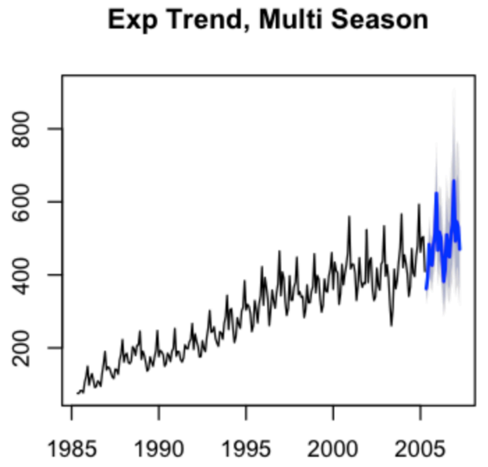
(For this next problem, important to note that trend=additive, exponential=FALSE are default and trend switches to multiplicative when exponential is switched to TRUE. I am just being explicit to be clear.)

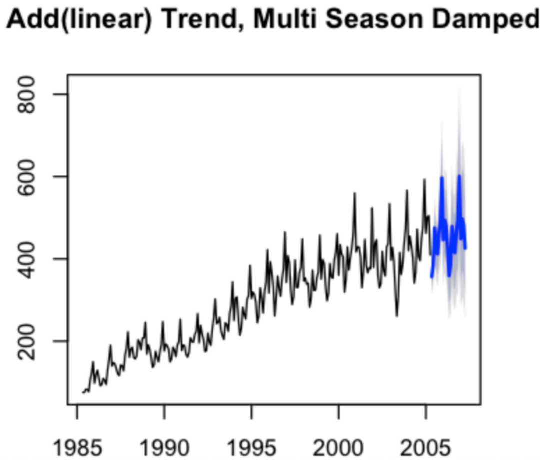
3.4. Model error declines for the plots below (annotated by UL UR BR BL):

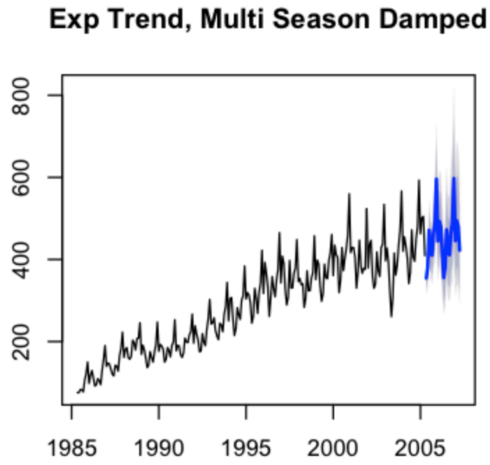
a. UL: RMSE = 14.66 for seasonal=mult, trend=add, expo=FALSE
b. UR: RMSE = 14.62 for seasonal=mult, trend=multip, expo=TRUE
c. BR: RMSE = 14.42 for* seasonal=mult*, dampened, trend=multip, expo=TRUE
d. BL: RMSE = 14.41 for seasonal=mult, dampened, trend=additive, expo=FALSE

I think these multiplicative trend models have lower RMSE because the data used to calculate RMSE includes and is dominated by nearly 15 years of strong trend upon which a multiplicative trend model would fit better than additive. I would stick with these models as they carry more information when calculating the trend and may prove useful for error generation as level increases. I believe these come at additional cost of complexity and possibly overfitting the future. Dampening seems similar to regularization via phi. Thus, I would select model (c) above. RMSE is not the lowest but I believe dampening assists in keeping the trend element from going too far.

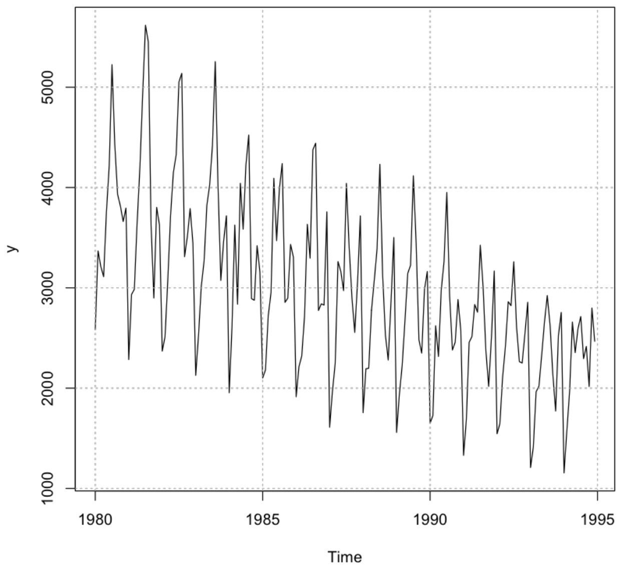

Add(linear) Trend, Multi Season


Exp Trend, Multi Season

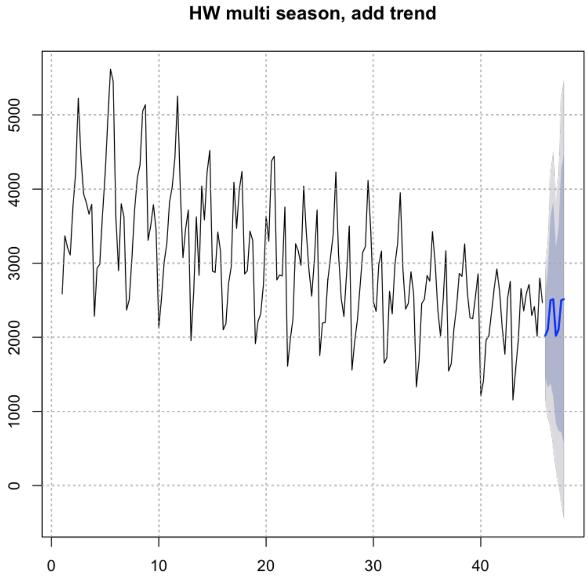

Add(linear) Trend, Multi Season Damped


Exp Trend, Multi Season Damped

4.1 Below is a plot. Principal features of this series is a strong negative trend that finished 2 years ago and which had significant but declining annual swings: The level has fallen from ~4000 mean annual value in Jan 1980 to ~2000 through about 1993. Variation has declined proportionately with level from variations of 3000 per year to less than 2000. In the last 2 years, variation has compressed significantly to ~750 and the level is in the top end of the 2-year range. It is noteworthy that the starting bottles and finishing bottles are roughly similar ~2500. Lastly in term of seasonal variation, over 15 years one can count 15 spikes up and 15 spikes down with some intra year variation since 1983 that is about half the length of those spikes up and down. The spikes from high to low have exceeded the mean level of wine sales in some circumstances and are therefore significant.

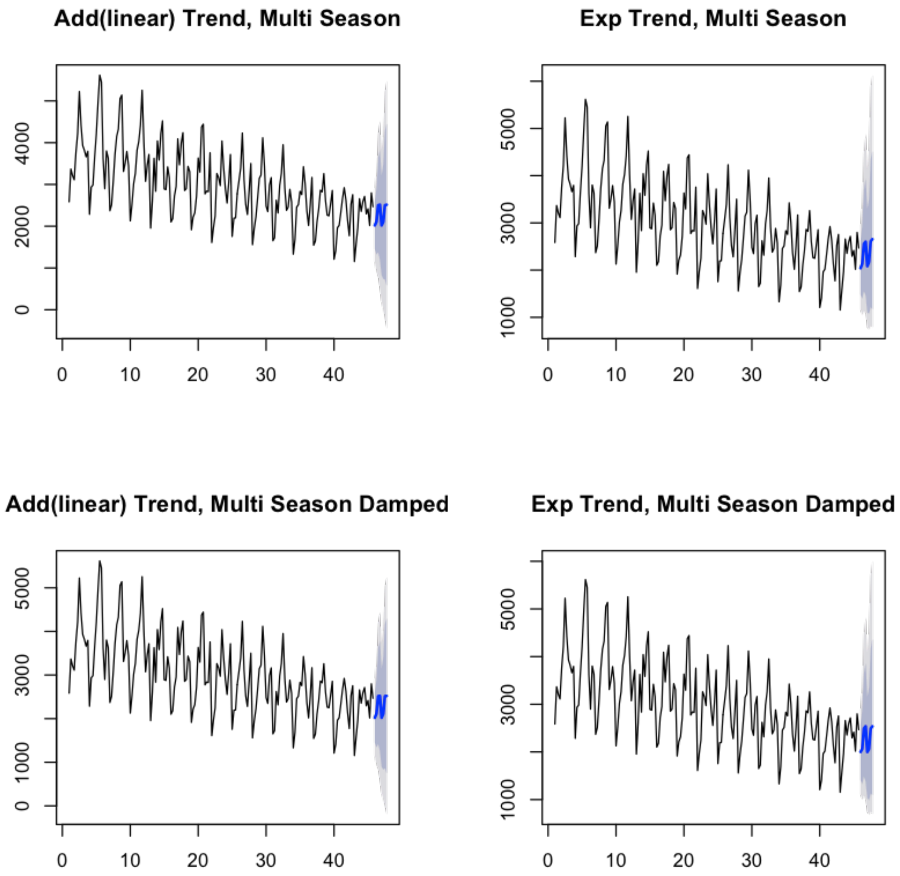


4.2 Below is a plot of HW multiplicative seasonality (defaults = additive trend / exponential=FALSE)



4.3 The reasons for using multiplicative seasonality (above) in this case is the same as in the case of Australia visitor arrivals. There is a trend which is reducing the level and we would want our seasonal moves to decrease by similar proportions. If not, the swings of the early 1980s would be multiple of the swings we have had now. I guess in some ways this enables us to retain information from the patterns from the 1980s without retaining their scale when producing forecasts.

4.4 If I am not mistaken, it appears that with additive trend and exponential = FALSE default settings, (labelled "Add(linear) Trend") in the 2 LHS plots below, the **forecast intervals are symmetric**, whereas for multiplicative trend, exponential TRUE (labelled "Exp Trend") in the 2 RHS plots below, the **forecast intervals are skewed to the higher end** of wine sales. This is likely because the trend of mean sales in the last 2 years was higher from ~1750 to ~2100 and the exponential multiplicative trend models (RHS) are picking up the direction and difference where the additive trend models (LHS) are picking up only the difference. There is not much of a difference between the additive trend (UL) and additive trend with dampening (BL) as the additive trend appears to be more of a local or very locally derived forecast, picking up only the last year of data and thus sideways. Dampening would not have an impact on a sideways forecast. On the other hand, the multiplicative trend (UR) has picked up some of the last 2 years' move higher and thus, the dampened multiplicative trend (BR) has tamped that down.


Add(linear) Trend, Multi Season

Exp Trend, Multi Season

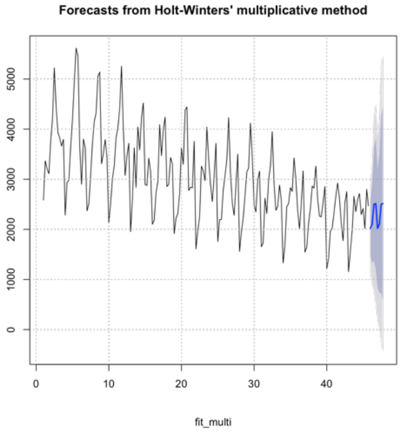
Add(linear) Trend, Multi Season Damped

Exp Trend, Multi Season Damped

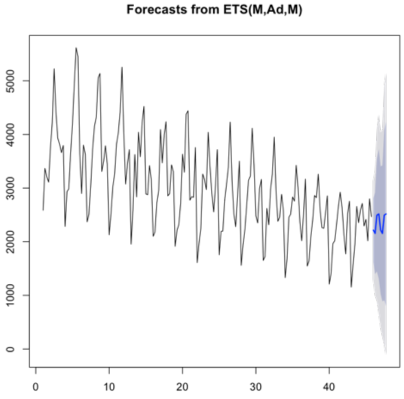
4.5 Comparing the RMSE from the 4 models above I prefer the same model as preferred for Australian visitors for the same reasons. Thus, **I prefer (c) the Exponential Multiplicative Damped Trend Multiplicative Seasonality model the most.** First, I like the fact that the forecast interval is skewed up slightly. I don't see the point including zero in such a wide forecast interval. I do not have a good statistical reason for my preference, but it seems unreasonable given the mere gradual trend lower, the tightening of variation, and the recent step up in the mean number of wine sales. The overall differences between RMSE amongst the 4 models is modest and the difference between the 2 multiplicative trend models is slight. I cannot recommend an additive trend model when trend has been a major feature in this series, locally for the last 2 years and globally over the last 15 years. The one I prefer (c) has the lowest RMSE and I feel the dampening effect might enable it to generalize better in validation and test, by dampening trend.

a. UL: RMSE = 580.3977 for seasonal=mult, trend=add, expo=FALSE
b. UR: RMSE = 570.5668 for seasonal=mult, trend=multipl, expo=TRUE
c. BR: RMSE **= 568.5359** ***for*** **seasonal=mult,** ***dampened, trend=multipl, expo=TRUE***
d. BL: RMSE = *571.4047* for seasonal=mult, dampened, trend=add, expo=FALSE

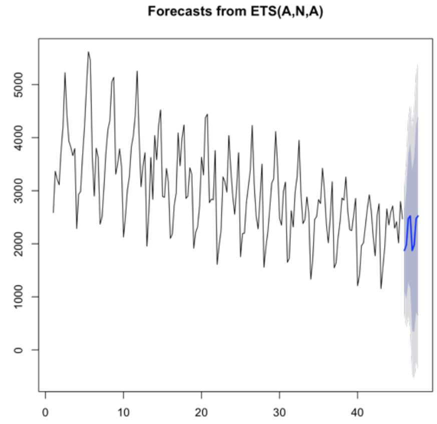
4.6.1 multiplicative HW method: Additive trend, Multiplicative Seasons



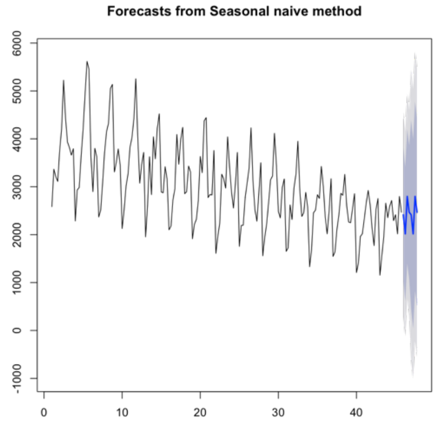
4.6.2 ETS: Multiplicative error, Damped Additive Trend, Multiplicative Seasonal



4.6.3 ETS: Additive error, No trend, Additive seasonality



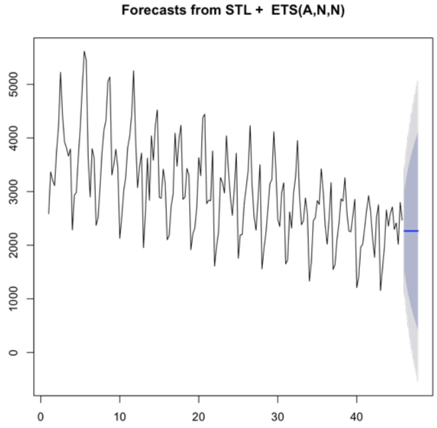
4.6.4 Naïve Seasonal



<u>Comment on 4.6.1 – 4.6.5</u>

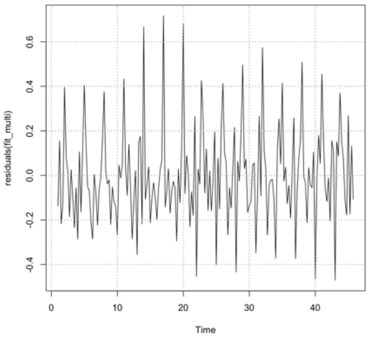
Of these 5 plots, the Naïve Seasonal in 4.6.4 is probably most interesting simply because the plot appears most realistic in directly copying the last year = last season of 4 quarters to the next 2 years forecast. 4.6.5 has no seasonal or trend component to modeling and simply broadcasts forward the detrended, deseasonalized point plus additive error. 4.61 and 4.62 employ additive trends, multiplicative seasonals. In trending series, I prefer multiplicative trends, but additive trends in these last 2 years of sideways movement presents a simple model that generalizes well.

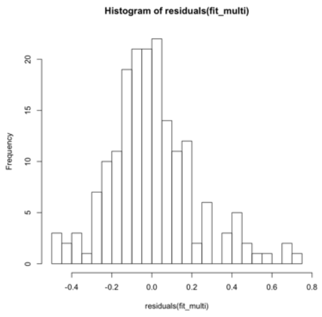
4.6.5 STL decomposition applied to Box Cox transformed data
+ ETS model (additive error, no trend, no seasonal)
applied to seasonally adjusted (transformed) data



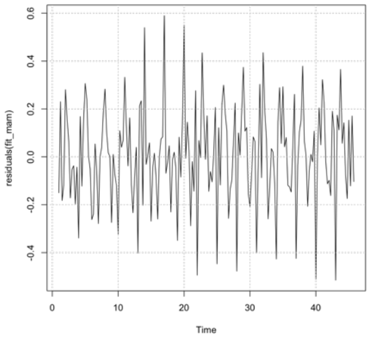
4.7 Residual diagnostics will number plots in the same way as above (replacing 4.6 with 4.7):

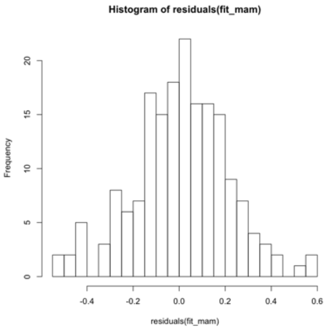
4.7.1 multiplicative HW method: Additive trend, Multiplicative Seasons



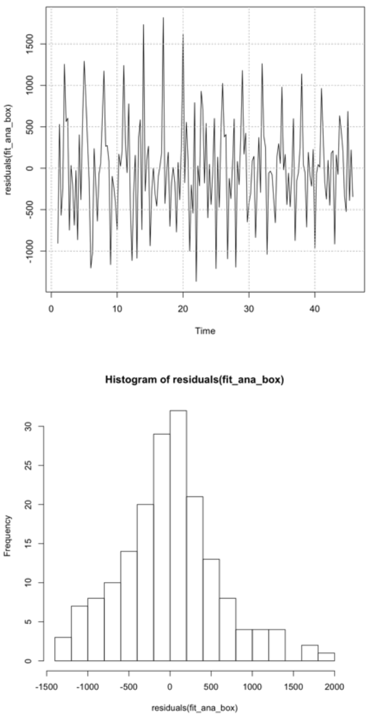


4.7.2 ETS: Multiplicative error, Damped Additive Trend, Multiplicative Seasonal

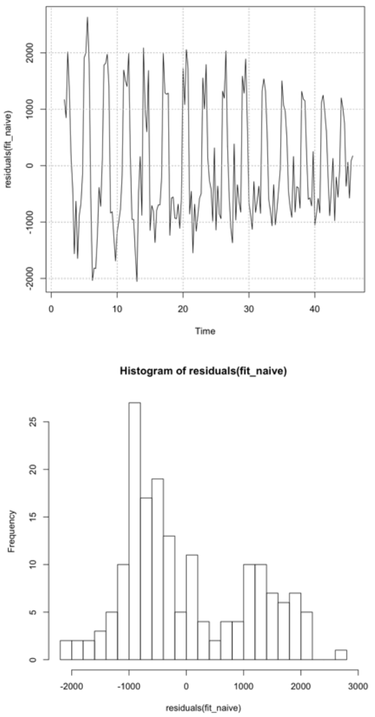




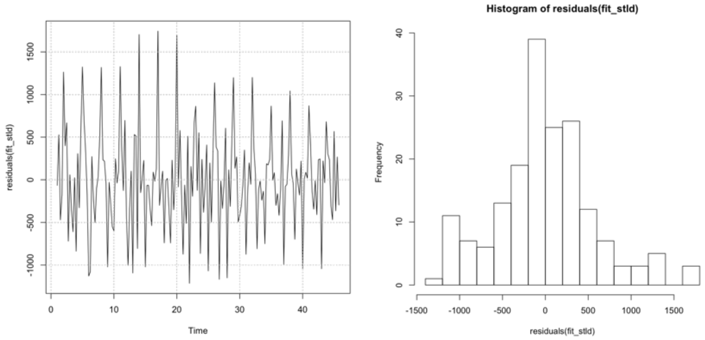
4.7.3 ETS: Additive error, No trend, Additive seasonality



4.7.4 Naïve Seasonal



4.7.5 STL decomposition applied to Box Cox transformed data
+ ETS model (additive error, no trend, no seasonal)
applied to seasonally adjusted (transformed) data



I solidly prefer models with normal residuals simply because they will likely generalize better in validation and subsequently when tested in prediction. This points squarely at 4.7.2 the **Multiplicative Error, Additive Trend, Multiplicative Seasonal model as my favorite**. As Hyndman states:

“The term ‘innovations’ comes from the fact that **all equations use the same random error process**, εt. For the same reason, this formulation is also referred to as a ‘single source of error’ model.”

I would prefer the single source of error to be a normally distributed residual. Other histograms skew right except 4.7.5 the de-seasonalized Box-Cox transformed with additive error, and 4.7.3 which also has additive error, no trend, and no seasonality.

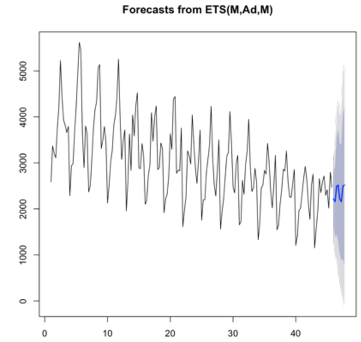
In terms of the temporal distribution of residuals, the Naïve model residuals compress moving LHS to RHS which makes sense since this model is very local. The Naïve in 4.74 has RMSE 1089.1 which is nearly double all others ranging 580-595. Most other models have residual compression as the plot moves LHS to RHS simply because alpha for most of these models has solved to between 0.83 and 0.86 and so are quite localized. 4.7.1 and 4.7.2 have similar model settings: both employ additive trend, multiplicative seasons, except 4.72 trends are damped employing a phi of 0.96 which is very conservative and probably when avoiding overfitting.

Lastly, I prefer multiplicative error models because their prediction interval may be skewed where for additive error models, the prediction interval is kept normal with median = point estimate = means. This seems too restrictive to me as mentioned when discussing the skewed forecast intervals in problem 4.4.

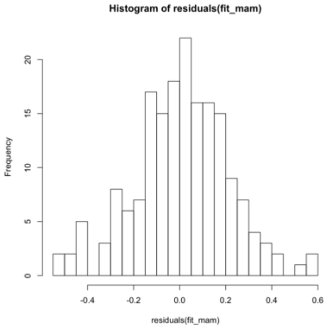
5.0

<u>Executive Summary</u>

After careful review of 15 years' fortified wine sales data and diverse set of 5 forecast models, I recommend employing a simple model that fits recently more stable sales data, which generalizes very well to capture seasonal variability, and which affords realistic dynamic forecast intervals we may use to gauge top line risk. Two years forecast sales data together with 80 and 95% forecast confidence levels are shown here:



Some specifics: This additive trend model is less complex and incorporates a 95% dampening factor. These features mitigate overfitting current sales data and overrunning future trends. Models of this kind have one random error process to derive optimal smoothing, trend, seasonal fits; of all models considered, this process generates residuals most representative of normal distribution for the spectrum of the last 15 years sales data. This suggest that the process does not leave any trend or seasonal relationship un modeled and that in the future it will apply equally well at all levels of sales we've experienced. Here's a plot of those residuals.


Histogram of residuals(fit_mam)
Frequency
residuals(fit_mam)